

# A Methodology for Anatomic Ultrasound Image Diagnostic Quality Assessment

**Hemmsen, Martin Christian; Lange, Theis; Brandt, Andreas Hjelm; Nielsen, Michael Bachmann; Jensen, Jørgen Arendt**

*Published in:*
I E E E Transactions on Ultrasonics, Ferroelectrics and Frequency Control

*Link to article, DOI:*
10.1109/TUFFC.2016.2639071

*Publication date:*
2017

*Document Version*
Peer reviewed version

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TUFFC.2016.2639071, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control



# A Methodology for Anatomic Ultrasound Image Diagnostic Quality Assessment

Martin Christian Hemmsen, Theis Lange, Andreas Hjelm Brandt, Michael Bachmann Nielsen and Jørgen Arendt Jensen

***Abstract*—This paper discusses methods for assessment of ultrasound image quality based on our experiences with evaluating new methods for anatomic imaging. It presents a methodology to ensure a fair assessment between competing imaging methods using clinically relevant evaluations. The methodology is valuable in the continuing process of method optimization and guided development of new imaging methods. It includes a three phased study plan covering from initial prototype development to clinical assessment. Recommendations to the clinical assessment protocol, software, and statistical analysis are presented. Earlier uses of the methodology has shown that it ensures validity of the assessment, as it separates the influences between developer, investigator, and assessor once a research protocol has been established. This separation reduces confounding influences on the result from the developer to properly reveal the clinical value. The paper exemplifies the methodology using recent studies of Synthetic Aperture Sequential Beamforming tissue harmonic imaging.**

## I. Introduction

NEW and improved imaging schemes are continuously being introduced in medical ultrasound for visualizing the anatomy. This includes improved B-mode schemes like non-linear imaging [1], synthetic aperture imaging [2], [3] and the derived method synthetic aperture sequential beamforming (SASB) [4], plane wave imaging [5], and minimum variance beamforming [6], [7]. Often these methods claim to improve on image quality with an underlying assumption that this translates into better diagnostic accuracy. To substantiate this claim, the new imaging method should be realistically compared to current imaging system in a clinical setting. The purpose of this paper is to give a fairly general approach to evaluating new imaging schemes for visualizing the anatomy based on our experiences with evaluating new synthetic aperture (SA) methods. Parts of the approach can also be used for investigations of e.g. flow imaging as performed in [8].

There are several tasks, which involve assessment of image quality. Equipment purchasing is partly based on performance specifications, acceptance testing verifies that the system fulfills the specified performance, quality assurance is used to ensure a constant system performance, clinical testing concentrates on the fulfillment of clinical needs, and optimization attempts to find best ways to use the imaging system for clinical purposes. These different tasks are best performed by different assessment methods, and the outcome is often referred to as technical (or physical) image quality or clinical image quality.

The term technical image quality is devoted to the direct measurable aspects of the image, and has for a long time mainly been focusing on spatial resolution, contrast, penetration depth, and uniformity. [9]–[15]. Historically in particular the spatial resolution has been used to characterize the imaging performance. However, this imaging performance metric can be misleading, and it is well recognized that the Rayleigh resolution or full-width-at-half-maximum (FWHM) criterion can not stand alone. Recommendations for measuring the comparative performance of medical ultrasound imaging equipment includes determination of the visibility of voids in a continuous background [16]–[18]. Imaging phantoms, however, does not provide a way to theoretically assess the performance of different hypothetical imaging systems. Although repeated image simulations can be performed to assess a wide variety of system parameters, this approach is very computationally challenging. Cystic resolution, a performance metric, conceptually related to the void-visibility, were suggested to help predicting the performance of new methods [19], [20]. It quantifies performance as the size of a void that produce a given contrast. The metric is useful in that it enables a straightforward optimization of parameters that affects image quality.

Researchers of ultrasound imaging methods are interested in assessing the quality of their methods to increase its performance and clinical usefulness. When new imaging methods are developed they are often favorably evaluated by phantom setups against a reference method. The evaluations often leave the question of method optimization open, and as a consequence, often conclude that the novel method is favorable. However, this is a far too wide conclusion as the developer is not separated from the investigator. The value of the evaluation is further questionably, because despite a massive effort by several groups, there is no international consensus about a complete protocol for technical image quality assessment. Furthermore, establishing the link between physical image quality measures and clinical utility has been pursued for decades, yet the relationship between the results of technical performance and clinical usability is not fully understood.

When one speaks of clinical image quality, the actual point of view and the definition of image quality are often left unspecified. In medical ultrasound, images are used to diagnose patients (diagnostic imaging) or to treat them (interventional imaging). Therefore, image quality is most meaningfully de-

Martin C. Hemmsen and Jørgen Arendt Jensen are with the Center for Fast Ultrasound Imaging, Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark.

T. Lange is with the Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark.

Andreas Hjelm Brandt and Michael Bachmann Nielsen are with the Department of Radiology, Copenhagen University Hospital, Copenhagen, Denmark.

fined through the use of the images in accomplishing these tasks.

Clinicians and policymakers often distinguish between the efficacy and the effectiveness of an intervention. Efficacy trials (explanatory trials) determine whether an intervention produces the expected result under ideal circumstances. Effectiveness trials (pragmatic trials) measure the degree of beneficial effect under real world clinical settings [21].

Fryback and Thornbury [22] presented a general six-tiered hierarchical model, which extends from basic laws of physics of imaging, through clinical use in decisions about diagnosis and intervention, to patient outcome and societal issues. The model can be used to classify assessment studies and provides a structure in which to relate efficacy to technology assessment and outcome research [23]. For an imaging examination to be efficacious at a higher level, it must be efficacious at lower levels; however, the reverse is not necessarily true. Increase in efficacy at lower level (e.g., improved imaging technical quality, level 1) does not guarantee improvement in efficacy at a higher level. The model is interesting, because it helps understand the importance of assessing the image quality at different levels.

The purpose of this paper is to disseminate best practices in relation to the assessment of image quality for new ultrasound imaging methods based on our experience of such assessment in a number of studies [24]–[26]. We have here been confronted by the many choices in making a fair comparison and ensure a statistically valid result. The paper presents a methodology that will potentially ensure a fair assessment between competing imaging methods, and is described in Section III. The methodology is inspired by the six-tiered hierarchical model and selected levels of the model is explained in Section II. The methodology suggests a three phased study plan covering the levels from technical assessment in the early prototype development phase, to the first initial preclinical trial and finally a clinical trial assessing the diagnostic accuracy and diagnostic thinking during the second and third phase. Based on recommendation 500 from the International Telecommunication Union - Radio-communication (ITU-R) for subjective quality assessment [27], a clinical assessment protocol and software is developed and presented in Section IV, and Section V presents the statistical analysis. The methodology is exemplified using a recent study of Synthetic Aperture Sequential Beamforming harmonic imaging.

## II. A Hierarchical Model of Efficacy

The hierarchical model is useful in discriminating between different levels of efficacy [22]; Table I gives a short overview. The following sections will go into more details about the first three levels. These levels are the relevant levels in relation to method optimization and guided development of new methods.

### A. Technical Efficacy

At the foundation of the hierarchy is assessment of technical efficacy: studies that are designed to determine if a particular proposed imaging method has the underlying ability to produce an image that contains useful information. Technical efficacy is generally the purview of developers concerned with the physical parameters. These include spatial resolution, contrast-to-noise ratio, signal-to-noise ratio, clutter-to-tissue ratio, speckle signal-to-noise ratio, uniformity, cystic resolution [19], [20], penetration depth, contrast detectability [29] among others. These parameters are usually derived under optimal laboratory conditions and only quantifies part of the image quality. Additional important variables include the presence of artifacts from the imaging itself. Determination of the technical efficacy is a prerequisite for consideration of efficacy at subsequent levels.

TABLE I
Hierarchical Model [a]

| |
|---|
| **Technical efficacy: production of an image or information** |
| Measures: signal-to-noise ratio, resolution, contrast, penetration, etc. |
| **Diagnostic accuracy efficacy: ability to differentiate between disease and nondisease.** |
| Measures: sensitivity, specificity, receive operator characteristic curves. |
| **Diagnostic thinking efficacy: impact on likelihood of diagnosis.** |
| Measures: diagnostic certainty. |
| **Treatment efficacy: potential to change therapy for a patient** |
| Measures: treatment plan, operative or medical treatment frequency. |
| **Outcome efficacy: effect on patient health** |
| Measures: mortality, quality-adjusted life years, health status. |
| **Societal efficacy: appropriateness from perspective of society.** |
| Measures: cost-effectiveness. |

[a] adapted from [28].

### B. Diagnostic-Accuracy Efficacy

The second level in the hierarchy determines if the imaging method predicts the truth. The study of accuracy is characterized by the attempt to measure performance for the purpose of making diagnoses and requires interpretation of an image by an observer. To determine the diagnostic-accuracy the true disease status of every subject is assumed to be known with certainty, either by an existing gold standard for indication of presence of such a disease/abnormality or by an independent assessment. Simple measures such as counting the number of abnormal patients found in a case series, sensitivity, and specificity are often used, but more sophisticated concepts of test performance, such as the receiver operating characteristic (ROC), are becoming more prevalent.

Receiver operating characteristic analysis has become a favored method for reflecting diagnostic accuracy [30]–[33], although there are acknowledged problems of spectrum bias in control patients [34], and problems of establishing the true (or gold standard) diagnosis [35]. In an ROC experiment each evaluated image or movie is assigned a rating, and by convention a higher rating indicates greater evidence of the presence of an abnormality.

Important to all measures of diagnostic-accuracy are that they attempt to measure performance of the imaging for the purpose of making diagnoses, and that they all require interpretation of the image by an observer, such as a radiologist. As such, diagnostic accuracy is the result from the joint function of image technical quality and interpretation by a observer.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TUFFC.2016.2639071, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control



Another important aspect to consider is that diagnostic accuracy is also a function of the physician requesting the examination, because the physician selects which patients will be imaged. As sensitivity and specificity can vary, depending on the spectrum of patients selected for imaging, this selection process can affect the result. Design of the clinical research, thus, requires a detailed study protocol to ensure both internal- and external validity [36].

### C. Diagnostic-Thinking Efficacy

Image information may change the physicians diagnostic certainty, change the differential diagnosis, strengthen a competing diagnostic hypothesis, or simply reassure the physician that no occult and unexpected is present. As such, at the third level the effect on the physicians certainty of a given diagnosis is evaluated.

The assessment of diagnostic thinking efficacy is relevant; because if the level of image quality is extremely low, the image provides little information for the diagnosis, and diagnostic accuracy is poor. When the image quality improves, important patterns become recognizable and diagnostic performance improves. But beyond a certain level, where the important features are already visible, and no additional image information that would be useful for the radiologist can be brought in the image, the diagnostic thinking performance will saturate. As such, even if there is a difference between methods in diagnostic-accuracy, they can provide the same level of diagnostic thinking efficacy.

Assessment of the diagnostic thinking efficacy requires no ground truth and can be measured by, for example, the difference in the clinician's certainty of a diagnose. For competing imaging methods one can measure whether the physician has relative greater or less confidence in the diagnosis using the new method.

## III. Methodology

The main issue in performing a structured and fair comparison between imaging methods, is to keep factors, such as transducer, scanner, region of interest, frame rate, and recording time constant. Other issues to consider is to get sufficient number of scans under realistic operating conditions and separating the developer and assessor in the evaluation process to remove personal bias. To fulfill these demands we propose that evaluations of new methods is conducted in a three stage research, as illustrated in Fig. 1:

The suggested methodology describes the research steps and experiments needed in an attempt to establish evidence of image quality differences between competing methods for ultrasound images depicting the anatomy. The methodology encompasses three phases, from demonstration of prototype to clinical assessment. The process of assessing a new method is beneficially split into three phases to allow a clear communicative process and to set focus on the different activities through the process. The three phases are performed in a sequential order, and if the assessment in one phase does not show evidence of improved image quality the following phases are skipped. The first phase is iterative in nature, and once a set of optimal parameters are determined, the setup is fixed and the next trial stage is begun.

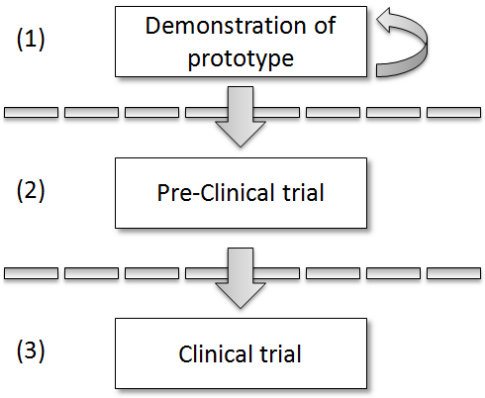


Fig. 1. Diagram of the methodology. Phase I relates to the demonstration of a prototype. Phase II relates to the exploratory investigation of the first trials conducted in accordance with the official standards (FDA or IEC) and approved by the local ethics committee. Phase III relates to the clinical trial that generates data on efficacy in a controlled clinical setting.

### A. Phase I: Demonstration of prototype

In this stage the developers demonstrate a workable prototype of the new imaging method with measurements on phantoms and a few in-vivo images. In a collaboration between the developer and ultrasound specialist, the new method's parameters are iteratively optimized to achieve the best possible setup. During the iterative process the technical efficacy is assessed. This demands that the method is at a stage of development, where the optimal set of parameters can be fixed during the investigation and tuning in Stage 1. The ultrasound specialist is allowed to change a defined set of imaging parameters during the scan to optimize image scans during all stages, and should be trained to perform this during Stage 1.

Assessment of technical efficacy is concerned with the physical parameters describing technical image quality. These are typically first estimated using simulation software such as Field II [37], [38] and determination of the point-spread-function (PSF) for linear imaging systems. For non-linear imaging more complicated simulation schemes like K-wave [39], Abersim [40], finite element methods [41], or a non-linear angular spectrum approach [42] can be used to also include effects distorting the PSF. Using the PSF the spatial resolution and cystic resolution [19], [20] are determined and used to optimize the method for a chosen clinical useful scenario. The parameters are usually derived under optimal conditions and needs to be validated using measurements on phantoms [10]. Once a suitable setup is determined and validated, the technical performance can be compared with that of a reference method. The reference method must be optimized with same care as the new method. Ideally the reference method is optimized by an independent and blinded developer or manufacturer determined parameters are used. A technical assessment in favor of the new method is a prerequisites for a consideration of assessment of efficacy at a higher level.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TUFFC.2016.2639071, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control



Before starting the clinical trials, measures are conducted to adjust the transmit levels such that intensities and temperature constraints are obeyed [43]–[46]. Once parameters of the two competing methods are fixed, a few in-vivo measurements are obtained to study the preliminary efficacy. Such tests compliments the technical assessment to decide whether the new method has scientific merit for further development and assessment. Furthermore, the in-vivo images helps the developer to decide which questions to answer using the clinical study.

### *B. Phase II: Pre-clinical study*

Phase II relates to the exploratory investigation of the first trials conducted in accordance with the United States Food and Drug Administration (FDA) or similar standards in other countries and approved by the local ethics committee. These trials are designed to establish early on whether the method behaves in human subjects as was expected from the prototype development. Typically the relevance of a clinical investigation is tested in a small group of people. It should be emphasized that pre-clinical here denotes a small trial on human volunteers prior to the real clinical trial, and is not a pre-clinical trial on animals.

The phase begins with the development of a clinical protocol [36]. The protocol describes carefully the planning and execution of the trial with clear objectives. The developed protocol describes the methods and its parameters in such a degree that the developer is and should be left out in the active part of the following research and should not have any influence on the outcome of the research in either data acquisition, any form of processing of it, or evaluation. The outcome of the trial is primarily the determination of feasibility, time requirement and cost of recruiting adequate numbers of eligible participants. The trial is also designed to demonstrate that planned measurements, data collection instruments and data management systems are feasible and efficient.

At this phase of the research it is important to be concise about which questions the developer wants to answer using the clinical trial. Also, the selection of test persons have an influence on the outcome of the clinical study. It is important to select representative test persons, such that the validity of the research questions are ensured. A key question is also to determine here how to present the data and how the assessor is educated to use the evaluation program. This phase ends when the outcome of the trial has been evaluated. If the outcome of the evaluation finds it feasible to conduct the clinical trial with the estimated number of participants (power calculation) and the equipment described in the clinical protocol, the clinical trial begins.

### *C. Phase III: Clinical study*

Clinical trials generate data on efficacy in a controlled clinical setting. When a study assesses efficacy, it is looking at whether the method is able to influence an outcome of interest (e.g. detection rate or diagnostic certainty) in the chosen population. At this stage of research the statistical significance of the new method is investigated by comparison to a reference method. Assessment of the method is performed by a number of ultrasound specialists independent to the method. Furthermore, the assessors must be separated from the specialists performing the ultrasound scanning, blinding them from the acquisition and any form of processing of it. The study is performed on a large enough group of subjects to test the hypothesis. The number of required subjects is determined in Phase II during the power calculation.

## IV. Assessment methodology

To date there exist no published methods that objectively assesses the clinical quality and efficacy of ultrasound images, and it is probably unlikely that a general, objective metric for this can be found, and as a consequence clinical image quality is assessed subjectively. One major limitation with subjective assessment is, if the opinion is just based on an impression of quality, the usefulness of the assessment is questionable. When judged by task-based criteria - for example by the opinion of the radiologist relating to his/her ability to recognize certain anatomical details or features in the image (diagnostic-accuracy) or his/her confidence on the perception of these details (diagnostic-thinking), the assessment is more relevant.

The proposed assessment methodology is based on earlier publications of studies of clinical evaluation between pairs of sequences [24]–[26] and suggested testing procedures according to recommendation 500 from ITU-R [27] for subjective quality assessment. The proposed methodology describes two assessment situations, one for the assessment of diagnostic-accuracy and one for diagnostic-thinking. A detailed description of a software toolbox to help in the assessment can be found in [47].

Assessors should be an expert observer, i.e. an observer that has expertise in the field of study and on image artifacts. Assessors should not be, or have been, directly involved, in the development of the system under study. The number of assessors needed depends upon the sensitivity and reliability of the test procedure and upon the anticipated size of the effect sought. This is similar in spirit to the power analysis preceding any Randomized Controlled Trial (RCT). In our setting there are, however, many more assumptions to be made before such a power analysis can be conducted. As a rule-of-thumb based on our studies we therefore suggest to use three assessors and 10 patients in Phase II. The Phase III design should be roughly double size in terms of assessors and patients per assessor depending on the efficacy found in phase II. If the study is very large or invasive we suggest to include a trained statistician to conduct a more formal sample size calculation.

Assessors should be carefully introduced to the method of assessment, the types of stimuli, the grading scale, the sequence, and timing. Training sequences demonstrating the range and the type of the stimuli to be assessed should be demonstrated in an introductory material, where the assessor is allowed to ask question to fully understand the task at hand.

A test session, see Fig. 2, should at most last up to one hour, and should be conducted in a darkened room. If the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TUFFC.2016.2639071, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control



assessor is not completed within one hour, a 15 min break is forced before continuing the assessment to minimize effects from tiredness. At the beginning of the first session, about five representative presentations should be introduced to stabilize the assessors opinion and expectations. The data issued from these presentations must not be taken into account in the results of the test. A random order should be used for both the introductory presentations and the final test. It is important that the sequence is arranged, so that any effects of tiredness or adaptation are balanced out from session to session.

A test session can either be focused on the assessment of diagnostic-accuracy or diagnostic-thinking.

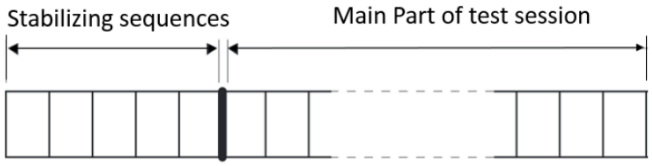


Fig. 2. Illustration of a test session. Introductory presentations are used to stabilize the assessors opinion and expectations. The main part of the test session is performed directly following the introductory presentations.

### A. Diagnostic Accuracy

The presentation method for assessment of diagnostic accuracy combines elements of the double stimulus continuous quality scale (DSCQS) method (ITU BT.500-11, Section 5) and the non-categorical judgment methods (ITU BT.500-11, Section 6.1.4.3). For reference, it may be called the sequential stimulus absolute scale (SSAS) method.

A test session comprises a number of presentations, each with a single observer. Unlike the DSCQS method where the assessor only observes the stimulus two times and rates each stimuli, the assessor is free to observe the stimuli until a decision is obtained.

Fig. 3a shows a basic test cell illustrating the presentation structure of reference and test material. Reference and test movies or images are displayed in a unique randomized sequential order. For movies, stimuli are visualized in a palindromic display fashion (looping forth and back). Fig. 3b shows the associated rating scale.

Fig. 3b illustrates the suggested rating scale. The scale is based on the non-categorical judgment method as described in ITU BT.500-11, Section 6.1.4.3. The judgment scale used is a numerical scale, where assessors assign a value to each stimuli that reflect the assessors certainty of abnormality. As such, the range of values are restricted to 0 to 100.

Test sessions consists of a series of test cells. These should be presented randomized, blinded, and independently of each other and, preferably, in a different random sequence for each observer. Table II illustrates the required stimuli for each test cell. Preferably, there would be at least 2 repetitions of each of the test cells to check for consistency.

Assessors are instructed to rate on a scale from 0 to 100% how certain they are that the stimuli contains an abnormality? They assess the sequence by placing a bar at the respective rating. Figure 4 illustrates the GUI associated with the rating process of diagnostic-accuracy.

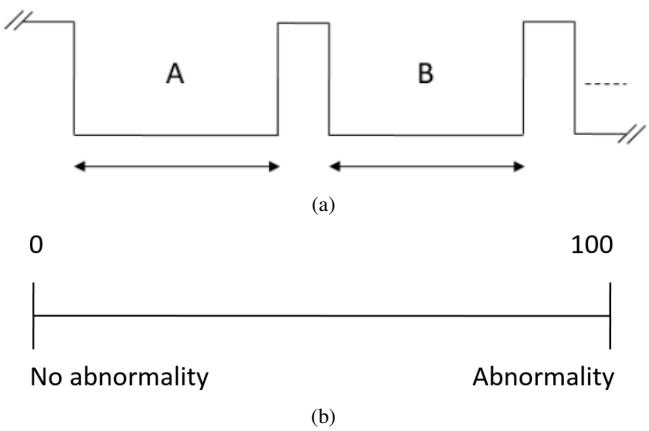


Fig. 3. Assessment of diagnostic-accuracy. (a) Basic test cell illustrating the presentation structure of reference and test material. Reference and test movies are displayed individual in randomized order. Here the reference stimuli is shown during time A and the stimuli from the new method is shown during time B. Assessors are free to observe the stimuli until a mental measure of certainty is obtained. (b) The rating scale used to quantify the certainty.

TABLE II
DESCRIPTION OF THE TEST CELL FOR ASSESSMENT OF DIAGNOSTIC-ACCURACY

| Stimuli |
|---|
| Reference sequence |
| Test sequence |

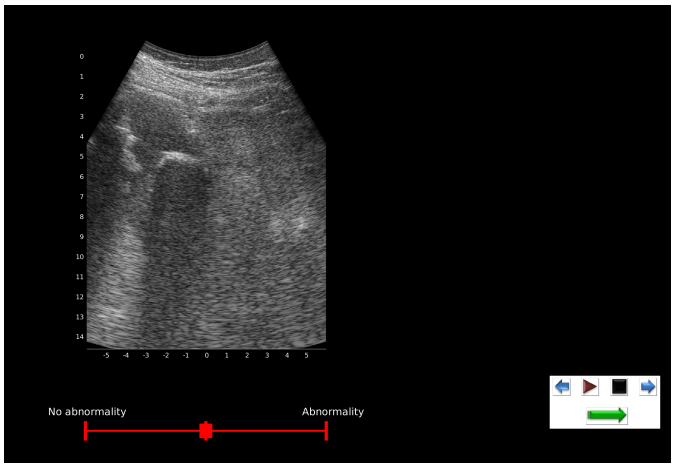


Fig. 4. Illustration of the GUI associated with the assessment of diagnostic-accuracy. The bar is placed at the respective rating where the assessor score their level of certainty that the stimuli contains an abnormality.

### B. Diagnostic Thinking

The presentation method for assessment of diagnostic certainty combines elements of the simultaneous double stimulus for continuous evaluation (SDSCE) method (ITU BT.500-11, Section 6.4) and the double stimulus continuous quality scale (DSCQS) method (ITU BT.500-11, Section 5). For reference, it may be called the simultaneous stimulus relative quality scale (SSRQS) method.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TUFFC.2016.2639071, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control



As with the SDSCE method, each trial will involve a split-screen presentation of material from two stimuli. One of the stimuli will be the reference, while the other is the test. The reference could be a conventional setup or the setup to compare against, and the test is the method under investigation. Unlike the SDSCE method, observers will be unaware of the scanner conditions represented by the two members of the stimuli pair and the left-right placements are randomized.

As with the DSCQS method, a test session comprises a number of presentations, each with a single observer. Unlike the DSCQS method, where the assessor only observes the stimulus two times and rates each stimuli, the assessor is free to observe the stimuli until a mental measure of relative quality associated with the stimulus is obtained. Fig. 5a shows a basic test cell illustrating the presentation structure of reference and test material. Reference and test stimuli are displayed as matching pairs side-by-side with random left-right placement. Stimuli are visualized in a palindromic (looping forth and back) display fashion to minimize discontinuity at the joints.

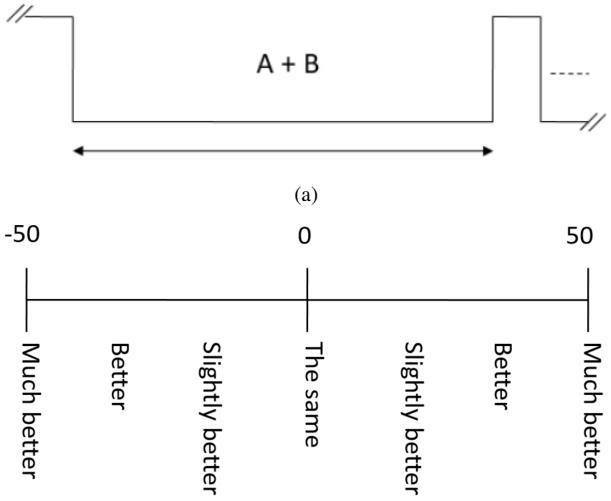


Fig. 5. Assessment of diagnostic certainty. (a) Basic test cell illustrating the presentation structure of reference and test material. Reference and test are displayed as matching pairs side-by-side with random left-right placement. Assessors are free to observe the stimuli until a mental measure of relative quality associated with the stimuli is obtained. (b) Visual analog scale (VAS) for diagnostic certainty comparison between left and right stimuli.

The most often used criteria for manufacturers to implement new processing methods in their equipment is better diagnostic value compared to the existing method. Accordingly, a stimulus comparison scale, as described in ITU BT.500-11, Section 6.2, is recommended to be used. The specific judgment scale used is a non-categorical (continuous) scale, as described in ITU BT.500-11, Section 6.2.4.2. For reference it may be called Visual Analog Scale (VAS). During introduction of the assessors to the system and the rating methods, VAS is described with the same number of labels as on the ITU-R categorical comparison scale, but with slightly modified labels (much better, better, slightly better, the same, slightly better, better, much better) to report the existence of perceptible quality differences and allow the random left-right placement of the stimuli. After introduction and during assessment the labels are hidden to avoid categorized data and to get a smoother distribution. Fig. 5b shows the associated VAS for diagnostic certainty comparison between left and right stimuli.

Test sessions consists of a series of test cells. These should be presented randomized, blinded, and independently of each other and, preferably, in a different random sequence for each observer. Table III illustrates the required stimuli for each test cell. Preferably, there would be at least 2 repetitions of each of the test cells to check for consistency. Note that each test cell consists of two pairs, the reference stimuli shown to the left and test stimuli to the right and vice verse. The two repetitions should not be displayed sequentially in time after each other, but randomized.

TABLE III
DESCRIPTION OF THE TEST CELL FOR ASSESSMENT OF DIAGNOSTIC-THINKING.

| Left stimuli | Right stimuli |
|---|---|
| Reference sequence | Test sequence |
| Test sequence | Reference sequence |

The judgment sessions should be divided into sittings not more than one hour in duration separated by a 15-minute rest periods. Assessors are instructed to evaluate whether they have a relative greater or less confidence in the diagnosis using the new method on a visual analog scale. Note here that it is not evaluated if the diagnose is correct, but merely if the observer feels more confident. As such the ground truth does not need to be known. Fig. 6 illustrates the GUI associated with the rating process of diagnostic-thinking.

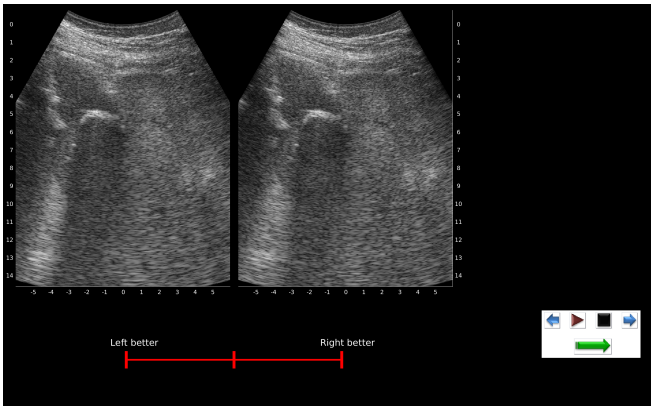


Fig. 6. Illustration of the GUI associated with the rating process of diagnostic thinking. Two stimuli are shown side-by-side, with the rating bar beneath.

## V. STATISTICAL ANALYSIS

The statistical analysis section is split into two sections, one for each clinical test.

### A. Diagnostic Accuracy

Assessing the diagnostic accuracy, assessors are instructed to rate on a scale from 0 to 100% how certain whey are that the stimuli contains an abnormality. In this two-class

prediction problem, the evaluated stimuli is assigned a rating and by convention a higher rating indicates greater evidence of the presence of an abnormality. A test is considered positive, if the rating exceeds a certain threshold $c$ representing the level of decision point. The agreement between a test and the true disease status can be summarized using two quantities: True Positive Ratio ($TPR$) and False Positive Ratio ($FPR$). $TPR$ is equivalent to sensitivity (Number of true positive assessment)/(Number of all positive assessment) and $FPR$ is equivalent to $1 - \text{specificity}$ (Number of false positive assessment)/(Number of all negative assessment).

A ROC curve is the plot of $TPF$ versus $FPF$, where the points on the graph are determined as the level of decision point, $c$, is varied. Thus, the ROC curve summarizes the agreement between ratings and the presence of an abnormality for all thresholds simultaneously. It is thereby simultaneously a tool to access the overall predictive quality, and a tool to determine the optimal cut-off point, $c$.

From [48] the ROC curve can be interpreted such that the faster the curve approach the upper left corner, the more useful the test results are. The slope of the tangent line to a cut-point tells us the ratio of the probability of identifying true positive over true negative, the likelihood ratio ($LR$). $LR = \text{sensitivity}/(1 - \text{specificity})$. If the ratio is equal to 1, the selected cut-point does not add additional information to identify true positive result. If the ratio is greater than 1, the selected cut-point help identify true positive result. If the ratio is less than 1, it decreases disease likelihood. The area under ROC curve ($AUC$) provides a way to measure the accuracy of a diagnostic test. The larger area, the more accurate the diagnostic test is. An area of 1 represents a perfect test; an area of 0.5 represents a worthless test. $AUC$ can be defined by the following equation, where $f = (1 - \text{specificity}) = FPR$ and $\text{ROC}(f)$ is sensitivity,

$$AUC = \int_0^1 \text{ROC}(f)\ \mathrm{d}f. \tag{1}$$

As the assessors might use the scales (the 0 to 100 score) differently a ROC curve with corresponding AUC estimate should be derived for each assessor and for each method separately. Standard software such as SPSS (the ROC-curve option in the Analyze menu) can produce ROC curves and AUC estimates along with associated standard errors directly from the provided scores. However, as the two methods and different assessors are depended (as they are all based on the same scans) formal statistical comparisons of methods cannot be straightforwardly done. It is possible to use the test for the AUC being above 0.5, which is the test for prediction accuracy above chance. In the next section we introduce a formal statistical test for the superiority of one method above the other.

### B. Diagnostic-Thinking

Intuitively a positive average VAS score will indicate that one method is superior to the other. However, due to the dependence between observations a standard $t$-test cannot be employed. Instead the VAS scores should be analyzed by a mixed effect linear model with a random effect [49] for each image pair and each assessor, thereby accounting for the dependence induced by repeatedly scoring the same image pair and collecting multiple scores from the same assessor. Since pairs are showed randomly either left or right, we do not have to take into account a possible preference for the image viewed on, say, the left side of the screen. Mathematically the mixed-effect model is given by

$$y_{i,j} = a_0 + \alpha_i + \beta_j + \epsilon_{i,j}, \tag{2}$$

where $y_{i,j}$ is the VAS score of the $j$'th assessor and the $i$'th scan, $\beta_j$ is a assessor specific random effect, $\alpha_i$ is a scan specific random effect, and $\epsilon_{i,j}$ is a measurement error. $\alpha$ and $\beta$ are assumed to follow potentially correlated mean zero normal distributions and $\epsilon$ an uncorrelated mean zero normal distribution. If the fixed parameter $a_0$ is significantly different from zero, there is evidence that one method is preferred. The model can be fitted in any statistical software package. In SPSS using the Repeated measures option in the Analyse menu. In R the nlme-package can be used. Model fit should be assessed by Q-Q plots [50] of the residuals from the mixed effect model.

## VI. Experiences from previous studies

The developed methodology has been used in a number of recent studies of Synthetic Aperture Sequential Beamforming (SASB). In this section, experiences from using the methodology and selected results are shown to demonstrate the use and work flow.

### A. Phase I - Demonstration of prototype

In this phase, parameters such as F#, focus distance, apodization etc. were optimized using simulation software and validated using measurements on a phantom with 2 wire targets. Table IV and Fig. 7 illustrates the result of the technical performance assessment. Conventional imaging using dynamic receive focusing (DRF) were chosen as the reference method, and was already implemented and optimized on an available commercial scanner (BK Ultrasound, ProFocus). The scanner allowed data acquisition of the two competing methods in an interleaved imaging mode, which ensured that data were acquired from the same anatomical region [47]. Parameters specified by the commercial manufacturer was used, and this ensured no conflict of interest during the image optimization.

TABLE IV
Result of the technical performance assessment. From left: Cystic resolution, Spatial resolution in lateral dimension, spatial resolution in axial dimension

| | $R_{12\text{dB}}$ [mm] | $\text{FWHM}_{\text{lat}}$ [mm] | $\text{FWHM}_{\text{ax}}$ [mm] |
|---|---|---|---|
| $\text{DRF}_{41\text{mm}}$ | 0.51 | 0.79 | 0.41 |
| $\text{SASB}_{41\text{mm}}$ | 0.46 | 0.71 | 0.41 |
| $\text{DRF}_{91\text{mm}}$ | 1.25 | 2.29 | 0.64 |
| $\text{SASB}_{91\text{mm}}$ | 1.69 | 1.54 | 0.56 |

The technical performance were supported with an in-vivo scan of a healthy volunteer, see Fig. 8. It was then concluded

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TUFFC.2016.2639071, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control



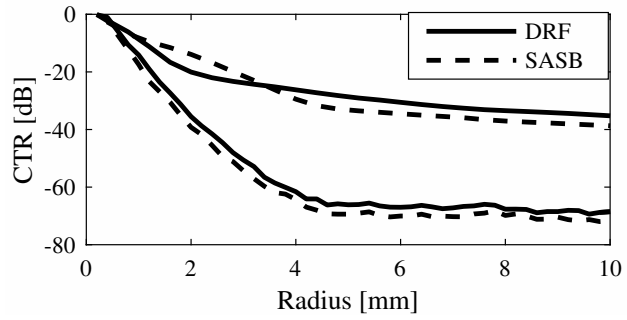


Fig. 7. Illustration of the cystic resolution as function of void size. DRF is shown using a black line and SASB using a dashed black line. The top graphs are at a depth of 91 mm and the bottom pair at 41 mm.

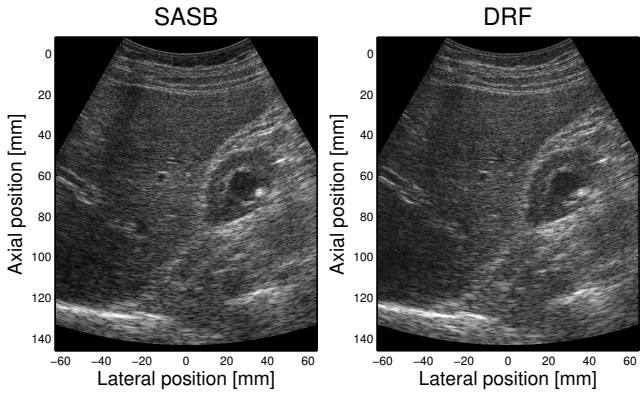


Fig. 8. Illustration of first in-vivo scan displaying the liver and tip of right kidney, (left) SASB and (right) Dynamic Receive Focusing.

that the method was ready for phase II. Parameters were locked and intensity measurements were performed. Hemmsen et al. [51] presents details of the study and its technical analysis.

### B. Phase II

In phase I the parameters and feasibility of the new method for abdominal imaging was investigated in a collaboration between developer and clinical specialists. In this phase, the developer no longer participate in the study and all acquisition and processing are performed by trained medical doctors. A study protocol was developed and approval from the local ethics committee was obtained. Initially 3 patients with malignant focal liver lesions (confirmed by biopsy or computed tomography/magnetic resonance) were included in the study. The patients were scanned in three positions, where the liver lesions were visible and in three areas where no pathology was visible. Data from the conventional technique and the new techniques, respectively dynamic receive focusing and SASB, were acquired simultaneously, giving images from the same anatomical location.

Using the described methodology for assessment of diagnostic accuracy and diagnostic thinking, two medical doctors (ultrasound specialists) evaluated the image sequences. None of the two were involved in the project, nor had they any prior knowledge about the details of SASB imaging, or seen any of the images beforehand. Evaluations were done blinded and independently of each other. Each sequence pair was first displayed two times with opposite left-right placement for assessment of diagnostic thinking efficacy. The sequence pairs were then split and the individual sequences were displayed for assessment of diagnostic accuracy. This gave 48 + 48 presentations of the 24 sequence pairs. Before the assessment the assessors were introduced to the rating program and instructed how to interpret the scales when performing the assessments. Labels on the scales were hidden during the actual evaluation to avoid categorized data. Before the actual assessment, five trial examples were shown to get the assessors acquainted with the task at hand and which types of images to expect.

Three medical doctors assessed the acquired stimuli and statistical analysis was performed. Fig. 9 illustrates the distribution of scores from the assessments of diagnostic-thinking efficacy. The details of the study and the statistical analysis, is presented in [52].

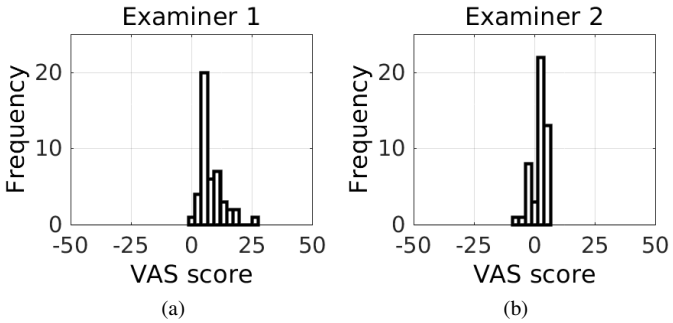


Fig. 9. Histograms from the assessment of diagnostic thinking. (a) assessor 1, (b) assessor 2. (From [52])

### C. Phase III

Based on the statistical analysis from phase II forty-three patients with different kinds of malignant focal liver cancer (primary liver tumor or liver metastasis) were asked to participate in the study. All patients were included after providing informed consent and on approval by the Danish National Committee on Biomedical Research Ethics (Journal No. H-1-2011-124). Before the study, liver lesions were diagnosed by biopsy or computed tomography/magnetic resonance (CT/MR). Before the experimental scan, an orientation scan was performed with a conventional ultrasound scanner (UltraView 800, BK Medical, Herlev, Denmark). Included were only patients in whom the pathology was visible on the orientation scan. Twelve patients were excluded because the pathology was not visible; thus, a total of 31 patients with focal liver cancer (28 colorectal liver metastases and 3 hepatocellular carcinomas) were examined with the experimental setup.

The patients were scanned in three positions, where the liver lesions were visible, and in three areas where no pathology was visible. The patients were positioned supine and were told to hold their breath and lie still during recording. The aim was to record six sequences for each patient, but because of technical challenges, this was possible for only 28 patients. One patient had only three recordings, and two patients had

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TUFFC.2016.2639071, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control



seven recordings because of errors made while saving and noticed after the scan session. A total of 185 image sequences were recorded. The recorded data were processed off-line with no user interaction. To ensure that clinically valuable image sequences were generated, a subsequent selection was performed before the assessments. Images defined as not clinically valuable were (i) sequences in which no liver tissue was visible, (ii) sequences in which malignant focal liver cancer was not visible even though it had been reported and (iii) sequences in which patient movement made the sequence impossible to assess. The selection was done blinded to knowledge of image technique.

Using the described methodology for assessment of diagnostic-accuracy and diagnostic-thinking, eight radiologists blinded to the methods assessed all image sequences. Each sequence pair was first displayed two times with opposite left-right placement for assessment of diagnostic thinking efficacy. The sequence pairs were then split and the individual sequences were displayed for assessment of diagnostic accuracy. This gave 254 + 254 presentations of the 127 sequence pairs. In total, 4,064 assessments were completed. The average duration to assess one set of image sequences was approximately one hour, i.e. 15 seconds on average per image sequence. The details of the study and the statistical analysis is presented in [26].

The ROC assessment was not conducted in [26], but the same data has been used by one medical doctor (AHB) to perform the assessment as an example, and the result is shown in Fig. 10. All images were reviewed and graded on a scale from 0 to 100 indicating the confidence of seeing a tumor in the image with 100 being absolute certainty. The results have then been converted into the ROC curve shown in Fig. 10. The area under the curve shows a slight advantage for SASB-THI compared to traditional non-linear imaging. This difference does not yield a clear clinical benefit of SASB, but it demonstrates that SASB images are similar in performance to traditional ultrasound images although a data reduction of a factor of 64 is attained by SASB. This demonstrates the possibility of introducing wireless probes without compromising image quality.

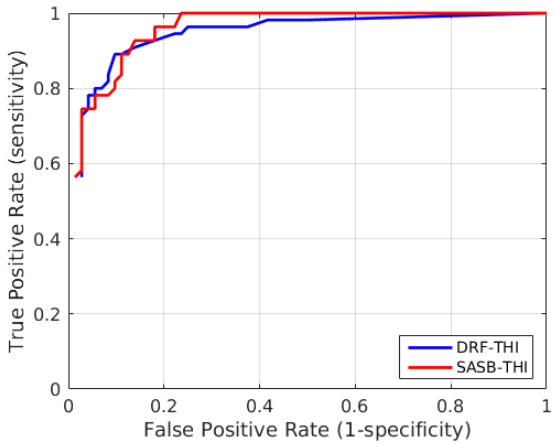


Fig. 10. Result of the diagnostic accuracy assessment for a single doctor.

## VII. Discussion and Conclusion

Based on the general six-tiered hierarchical model by Fryback and Thornbury, a three phased methodology for assessing the quality of new imaging methods has been presented and demonstrated. The three phases of the methodology describes an assessment of competing methods from basic laws of physics, through clinical use in decisions about diagnosis. Recommendations to the clinical assessment protocol, software, and statistical analysis are presented. Earlier uses of the methodology on linear and non-linear imaging using SA and SASB has shown that the methodology gives valid assessments [24]–[26], as it separates the developer, investigator, and assessor once a research protocol has been established. This separation eliminates confounding influence on the result from the developer, as the successful introduction of new ultrasound imaging methods is driven by their clinical value. The methodology was exemplified using recent studies of Synthetic Aperture Sequential Beamforming tissue harmonic imaging, and its development was based on a number of studies conducted by our group.

The developed framework is a starting point for designing clinical studies for comparing two imaging methods as it forces a structure on the process and exemplifies the decisions on how to conduct the study. The first stage ensures the completion of the method in that all choices have to be taken before it can be implemented in a commercial scanner. It should not be possible to change and tweak parameters during the clinical study, as this is not a realistic option in a real clinical setting. Some parameters should be changeable by the radiologist, but often it is difficult in an experimental set-up in real-time to have all options available as the final processing is not real-time. In many studies data collection must be separated from the final processing as the demands in e.g. SA imaging [24], minimum variance beamforming, or 3-D imaging are too demanding for a real time implementation. This can have a biasing effect, as the real time orientation image does not reveal the same details as the improved approach, and therefore the scan angles would be different. It is, thus, important that the radiologist performing scans is experienced to choose the right views. In this process there might also be limiting factors in having two imaging sequences being emitted inter-spaced. This will often reduce the frame rate by a factor of two [25] and it can also limit the energy possible to emit. Having a poor orientation image during data acquisition can also lead to poor image views, and such images should be removed from the evaluation, as they have no clinical values. The person removing these images should not be part of the evaluation, and the object criteria for removal should be clarified.

The evaluation is performed either by finding features in a single image or comparing two images side by side. Blinding and randomizing the evaluation takes care of many biasing effects, but there are still cases where tradition or other effect might be limiting. For new imaging methods like minimum variance [7], [53] or spatial coherence [54], [55] the images might appear radically different than conventional ultrasound images, and this will in general bias the evaluation towards

the conventional image and preclude new inventions. This is a general problem in a necessarily conservative field. Clinical significance as described in [56] is a metric similar to the ROC curve, but it can be more appropriate when studies report efficacy in terms of a continuous measurement. The evaluation can also be dependent on the experience of the evaluator, and this can be separated out in the statistical processing, or all evaluations can be polled to reflect the diversity of radiologist using ultrasound.

A vital point in any comparison is the reference image. It can be problematic to have the developer of new algorithms also optimize the reference method, as there is an inherent bias towards the new method. The best approach is therefore to have an independent reference in the form of a commercial image separately optimized. A further problem with commercial images if often the heavy post-processing intended to improve image quality and remove artifacts. This is a very important part of modern ultrasound imaging, but obscures the real benefit of new acquisition schemes. Two approach can reduce these effects: The processing can be disabled on both images. This ensures a fair comparison of the basic image acquisition and processing. The drawback is the ”raw” image quality, which often is unusual for the evaluators to judge. The second approach is to enable post-processing on both images. This has the drawback that the processing is optimized for the conventional image and not for the new image like phase coherence [54]. We have used the first approach of disabling post-processing in the studies we currently have conducted, and have often been confronted with the rather poor image quality of ”raw” images by the evaluators.

This also touches on the learning aspect of the evaluation. What should the assessor expect and how should the relative VAS scale be used? The selection of the training set and the presentation of this is difficult, as it affects the outcome, and it is a constant discussion during the planning of a study on how to select and present the training set, what questions to ask, and how to instruct evaluators [26].

There is no doubt that making an evaluation of new imaging methods is complicated and affected by many factors. The mere method of presenting a few random examples and interpret them by scientists involved in the research is insufficient. This paper have presented our ideas and thoughts about how to evaluate methods in a less biased and more quantitative way. It will not work for every situation, but it hopefully spurs some thoughts and discussion and forward the field of improving ultrasound imaging by introducing new methods valuable in the clinic.

## REFERENCES


[1] B. Ward, A. C. Baker, and V. F. Humphrey, “Nonlinear propagation applied to the improvement of resolution in diagnostoc medical ultrasound,” *J. Acoust. Soc. Am.*, vol. 101, no. 1, pp. 143–154, 1997.

[2] M. H. Bae and M. K. Jeong, “A study of synthetic-aperture imaging with virtual source elements in B-mode ultrasound imaging systems,” in *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 47, 2000, pp. 1510–1519.

[3] K. L. Gammelmark and J. A. Jensen, “Multielement synthetic transmit aperture imaging using temporal encoding,” *IEEE Trans. Med. Imag.*, vol. 22, no. 4, pp. 552–563, 2003.

[4] J. Kortbek, J. A. Jensen, and K. L. Gammelmark, “Sequential beamforming for synthetic aperture imaging,” *Ultrasonics*, vol. 53, no. 1, pp. 1–16, 2013.

[5] J. Y. Lu, “2D and 3D high frame rate imaging with limited diffraction beams,” *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 44, pp. 839–855, 1997.

[6] J.-F. Synnevåg, A. Austeng, and S. Holm, “Minimum Variance Adaptive Beamforming Applied to Medical Ultrasound Imaging,” in *Proc. IEEE Ultrason. Symp.*, vol. 2, Sept. 2005, pp. 1199–1202.

[7] I. K. Holfort, F. Gran, and J. A. Jensen, “Broadband Minimum Variance Beamforming for Medical Ultrasound Imaging,” *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 56, no. 2, pp. 314–325, 2009.

[8] K. L. Hansen, J. Udesen, N. Oddershede, L. Henze, C. Thomsen, J. A. Jensen, and M. B. Nielsen, “In vivo comparison of three ultrasound vector velocity techniques to MR phase contrast angiography,” *Ultrasonics*, vol. 49, pp. 659–667, 2009.

[9] AIUM, *The AIUM 100 mm test object and recommended procedures for its use*. Rockville, MD: American Institute of Ultrasound in Medicine, 1974.

[10] M. M. Goodsitt, P. L. Carson, S. Witt, D. L. Hykes, and J. M. Kofler, “Real-time B-mode ultrasound quality control test procedures,” *Med. Phys.*, vol. 25, no. 8, pp. 1385–1405, 1998.

[11] N. M. Gibson, N. J. Dudley, and K. Griffith, “A computerised quality control testing system for B-mode ultrasound,” *Ultrasound Med. Biol.*, vol. 27, no. 12, pp. 1697–1711, 2001.

[12] K. Brendel, L. Filipczynski, R. Gerstner, C. Hill, G. Kossoff, G. Quentin, J. Reid, J. Saneyoshi, J. Somer, A. Tchevnenko, and P. Wells, “Methods of measuring the performance of ultrasonic pulse-echo diagnostic equipment,” *Ultrasound Med. Biol.*, vol. 2, no. 4, pp. 343–350, 1977.

[13] E. L. Madsen, “Quality assurance for grey-scale imaging,” *Ultrasound Med. Biol.*, vol. 26, no. Suppl. 1, pp. S48–S50, 2000.

[14] J. M. Thijssen, G. Weijers, and C. L. de Korte, “Objective performance testing and quality assurance of medical ultrasound equipment,” *Ultrasound Med. Biol.*, vol. 33, no. 3, pp. 460–471, 2007.

[15] J. Satrapa, H. J. Schultz, and G. Doblhoff, “Automated quality control of ultrasonic B-mode scanners by applying an TMM 3D cyst phantom,” *Ultraschall in der Medizin*, vol. 27, no. 3, pp. 262–272, 2006.

[16] AIUM, “Technical standards committee. Standard methods for measuring performance of pulse-echo ultrasound imaging equipment,” American Institute of Ultrasound in Medicine, Bethesda, Maryland, Tech. Rep., 1990.

[17] ——, “Technical standard committee. Methods for measuring performance of pulse-echo ultrasound imaging equipment, part 2: Digital methods stage 1,” American Institute of Ultrasound in Medicine, Bethesda, Maryland, Tech. Rep., 1995.

[18] ——, “Technical standard committee. Quality assurance manual for gray-scale ultrasound scanners, stage 2,” American Institute of Ultrasound in Medicine, Bethesda, Maryland, Tech. Rep., 1995.

[19] D. Vilkomerson, J. Greenleaf, and V. Dutt, “Towards a Resolution Metric for Medical Ultrasound Imaging,” in *Proc. IEEE Ultrason. Symp.*, 1995, pp. 1405–1410.

[20] K. Ranganathan and W. F. Walker, “Cystic resolution: A performance metric for ultrasound imaging systems,” *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 54, no. 4, pp. 782–792, 2007.

[21] R. H. Brook and K. N. Lohr, “Efficacy, effectiveness, variations, and quality boundary-crossing research,” *Medical Care*, vol. 23, no. 5, pp. 710–722, 1985.

[22] D. G. Fryback and J. R. Thornbury, “The efficacy of diagnostic imaging,” *Medical Decision Making*, vol. 11, no. 2, pp. 88–94, 1991.

[23] C. M. Clancy and J. M. Eisenberg, “Outcomes research: Measuring the end results of health care,” *Science*, vol. 282, no. 5387, pp. 245–246, 1998.

[24] M. H. Pedersen, K. L. Gammelmark, and J. A. Jensen, “In-vivo evaluation of convex array synthetic aperture imaging,” *Ultrasound Med. Biol.*, vol. 33, pp. 37–47, 2007.

[25] M. C. Hemmsen, P. M. Hansen, T. Lange, J. M. Hansen, K. L. Hansen, M. B. Nielsen, and J. A. Jensen, “In vivo evaluation of synthetic aperture sequential beamforming,” *Ultrasound Med. Biol.*, vol. 38, no. 4, pp. 708–716, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301562911015729

[26] A. H. Brandt, M. C. Hemmsen, P. M. Hansen, S. S. Madsen, P. S. Krohn, T. Lange, K. L. Hansen, J. A. Jensen, and M. B. Nielsen, “Clinical evaluation of synthetic aperture harmonic imaging for scanning focal malignant liver lesions,” *Ultrasound in Medicine & Biology*, vol. 41, no. 9, pp. 2368–75, 2015.

[27] ITU, "Recommendation 500-11: Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Tech. Rep., 1974, 1974-2002.

[28] L. Medina, C. C. Blackmore, and K. Applegate, *Evidence-Based Imaging: Improving the Quality of Imaging in Patient Care*. Springer New York, 2011. [Online]. Available: https://books.google.dk/books?id=w-t6Ymr_Q6UC

[29] S. W. Smith, R. F. Wagner, J. M. Sandrik, and H. Lopez, "Low contrast detectability and contrast/detail analysis in medical ultrasound," *IEEE Trans. Son. Ultrason.*, vol. 30, no. 3, pp. 164–173, May 1983.

[30] C. E. Metz, "Special articles roc methodology in radiologic imaging," *Investigative Radiology*, vol. 21, no. 9, pp. 720–733, 1986.

[31] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285–1293, 1988.

[32] C. E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," *Investigative Radiology*, vol. 24, no. 3, pp. 234–245, 1989.

[33] J. A. Hanley, "Receiver operating characteristic (ROC) methodology: The state of the art," *Critical Reviews in Diagnostic Imaging*, vol. 29, no. 3, pp. 307–335, 1989.

[34] D. Gur, J. L. King, H. E. Rockette, C. A. Britton, F. L. Thaete, and R. J. Hoy, "Practical issues of experimental roc analysis: Selection of controls," *Investigative Radiology*, vol. 25, no. 5, pp. 583–586, 1990.

[35] R. M. Henkelman, I. Kay, and M. Bronskill, "Receiver operator characteristic (roc) analysis without truth," *Medical Decision Making*, vol. 10, no. 1, pp. 24–29, 1990.

[36] S. B. Hulley, *Designing clinical research*, ser. Designing Clinical Research. Lippincott Williams & Wilkins, 2007. [Online]. Available: https://books.google.dk/books?id=_7UWxJ5erSsC

[37] J. A. Jensen, "Field: A program for simulating ultrasound systems," *Med. Biol. Eng. Comp.*, vol. 10th Nordic-Baltic Conference on Biomedical Imaging, Vol. 4, Supplement 1, Part 1, pp. 351–353, 1996.

[38] J. A. Jensen and N. B. Svendsen, "Calculation of pressure fields from arbitrarily shaped, apodized, and excited ultrasound transducers," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 39, pp. 262–267, 1992.

[39] B. T. Cox and P. C. Beard, "Fast calculation of pulsed photoacoustic fields in fluids using k-space methods," *Journal of the Acoustical Society of America*, vol. 117, no. 6, pp. 3616–3627, 2005.

[40] J. P. Remenieras, O. B. Matar, V. Labat, and F. Patat, "Time-domain modeling of nonlinear distortion of pulsed finite amplitude sound beams," *Ultrasonics*, vol. 38, pp. 305–311, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0041624X99001122

[41] G. Pinton, G. Trahey, and J. Dahl, "Erratum: Sources of image degradation in fundamental and harmonic ultrasound imaging: a nonlinear, full-wave, simulation study [apr 11 754-765]," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 58, no. 6, pp. 1272–1283, June 2011.

[42] Y. Du and J. A. Jensen, "Fast simulation of non-linear pulsed ultrasound fields using an angular spectrum approach," *Ultrasonics*, vol. 53, no. 2, pp. 588–594, 2013.

[43] FDA, "Information for manufacturers seeking marketing clearance of diagnostic ultrasound systems and transducers," Center for Devices and Radiological Health, United States Food and Drug Administration, Tech. Rep., 2008.

[44] FDA, "510(k) guide for measuring and reporting acoustic output of diagnostic medical devices," Center for Devices and Radiological Health, FDA, Tech. Rep., 1985.

[45] J. Jensen, J. B. Olesen, M. B. Stuart, P. M. Hansen, M. B. Nielsen, and J. A. Jensen, "Vector velocity volume flow estimation: Sources of error and corrections applied for arteriovenous fistulas," *Ultrasonics*, vol. 70, pp. 136–146, 2016.

[46] J. A. Jensen, M. F. Rasmussen, M. J. Pihl, S. Holbek, C. A. Villagomez-Hoyos, D. P. Bradway, M. B. Stuart, and B. G. Tomov, "Safety assessment of advanced imaging sequences, I: Measurements," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 63, no. 1, pp. 110–119, 2016.

[47] M. C. Hemmsen, S. I. Nikolov, M. M. Pedersen, M. J. Pihl, M. S. Enevoldsen, J. M. Hansen, and J. A. Jensen, "Implementation of a versatile research data acquisition system using a commercially available medical ultrasound scanner," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 59, no. 7, pp. 1487–1499, 2012.

[48] W. Zhu, N. Zeng, N. Wang *et al.*, "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations," in *NESUG proceedings: health care and life sciences*, 2010, pp. 1–9.

[49] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics.*, vol. 38, no. 4, pp. 963–974, December 1982.

[50] M. B. Wilk and R. Gnanadesikan, "Probability plotting methods for the analysis of data," *Biometrika*, vol. 55, no. 1, pp. 1–17, 1968.

[51] M. C. Hemmsen, J. Rasmussen, and J. A. Jensen, "Tissue harmonic synthetic aperture ultrasound imaging," *J. Acoust. Soc. Am.*, vol. 136, no. 4, pp. 2050–2056, 2014.

[52] J. H. Rasmussen, M. C. Hemmsen, S. S. Madsen, P. M. Hansen, M. B. Nielsen, and J. A. Jensen, "Preliminary study of synthetic aperture tissue harmonic imaging on in-vivo data," in *Proc. SPIE Med. Imag.*, vol. 8675, 2013, pp. 1–10.

[53] J.-F. Synnevåg, A. Austeng, and S. Holm, "Adaptive Beamforming Applied to Medical Ultrasound Imaging," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 54, no. 8, pp. 1606–1613, Aug. 2007.

[54] J. Camacho, M. Parrilla, and C. Fritsch, "Phase coherence imaging," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 56, no. 5, pp. 958–974, 2009.

[55] M. A. Lediju, G. E. Trahey, B. C. Byram, and J. J. Dahl, "Short-lag spatial coherence of backscattered echoes: imaging characteristics," *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.*, vol. 58, no. 7, pp. 1377–1388, 2011.

[56] M. A. L. Bell, R. Goswami, J. Kisslo, J. Dahl, and G. Trahey, "Short-lag spatial coherence imaging of cardiac ultrasound data: Initial clinical results," *Ultrasound Med. Biol.*, vol. 39, no. 10, pp. 1861–1874, 2013.

**Martin Christian Hemmsen** earned his Master of Science degree in electrical engineering in 2008 and the Ph.D. degree in 2011, both from Technical University of Denmark (DTU). He has been an associate professor in biomedical engineering at the department of Electrical Engineering, DTU and now works in industry. His research centered around simulation of ultrasound imaging, synthetic aperture imaging, innovation of hand held ultrasound imaging systems, and image perception and quality assessment.

**Theis Lange** TL got his PhD in 2008 in mathematical statistics from the University of Copenhagen. His research involves theoretical and methodological work on statistics as well as a wide range of applied collaborations with medical doctors and epidemiologists (63 papers published within the last 5 years). His methodological work is focused on causal inference i.e. a formal framework for addressing cause and effect from observational data as well as trial data. His applied collaborations span from smaller RCT's to complex longitudinal observational studies. In 2012 he was awarded the Kenneth Rothman Prize for the best paper in Epidemiology. Currently TL holds a double affiliation as associate professor at the Section of Biostatistics, University of Copenhagen and visiting professor at Center for Statistical Science, Peking University.

**Andreas Hjelm Brandt** was born in Denmark in 1982. He earned his degree as a Medical Doctor at the Faculty of Health, Copenhagen University in Denmark 2010. He has been working at the Department of Radiology at Copenhagen University Hospital since 2012 and is pursuing his Ph.D. degree in radiology in collaboration with the Center for Fast Ultrasound Imaging at the Technical University of Denmark since 2014. His research includes in vivo vector flow ultrasound and image quality assessment.

**Michael Bachmann Nielsen** is a Medical graduate from the Faculty of Health Science, University of Copenhagen in 1985, and earned his Ph.d. degree in 1994 and doctoral degree (dr.med. dissertation) in 1998. He has published more than 180 peer reviewed journal articles on ultrasound or radiology. He currently holds a full professorship in Onco-radiology at the University of Copenhagen and a position as Consultant at the Department of Radiology, Rigshospitalet in Copenhagen. The research is centered around clinical testing of new ultrasound techniques, tumor vascularity assessed with CT perfusion techniques as well as training in ultrasound.

**Jørgen Arendt Jensen** earned his Master of Science in electrical engineering in 1985 and the PhD degree in 1989, both from the Technical University of Denmark. He received the Dr.Techn. degree from the university in 1996. He has since 1993 been full professor of Biomedical Signal Processing at the Technical University of Denmark at the Department of Electrical Engineering and head of Center for Fast Ultrasound Imaging since its inauguration in 1998. He has published more than 450 journal and conference papers on signal processing and medical ultrasound and the book ”Estimation of Blood Velocities Using Ultrasound”, Cambridge University Press in 1996. He is also the developer and maintainer of the Field II simulation program. He has been a visiting scientist at Duke University, Stanford University, and the University of Illinois at Urbana-Champaign. He was head of the Biomedical Engineering group from 2007 to 2010. In 2003 he was one of the founders of the biomedical engineering program in Medicine and Technology, which is a joint degree program between the Technical University of Denmark and the Faculty of Health and Medical Sciences at the University of Copenhagen. The degree is one of the most sought after engineering degrees in Denmark. He was chairman of the study board from 2003-2010 and adjunct professor at the University of Copenhagen from 2005-2010. He has given a number of short courses on simulation, synthetic aperture imaging, and flow estimation at international scientific conferences and teaches biomedical signal processing and medical imaging at the Technical University of Denmark. He has given more than 60 invited talks at international meetings, received several awards for his research, and is an IEEE Fellow. His research is centered around simulation of ultrasound imaging, synthetic aperture imaging, vector blood flow estimation, and construction of ultrasound research systems.